

# A new variational family for Bayesian phylogenetics

**Lloyd T. Elliott<sup>1</sup>, Evan Sidrow<sup>1</sup>, Alexandre Bouchard-Côté<sup>2</sup>**

<sup>1</sup>Simon Fraser University, <sup>2</sup>University of British Columbia

EcoSta, Tokyo, August 2025

# Genetics

- COVID-19 RNA genome:
  - AUUAAAGGUUUAUACCUUCC ...
- Human DNA genome:
  - TAACCCTAACCCTAACCCTA ...

SNP: Single Nucleotide Polymorphism (basepair substitution)

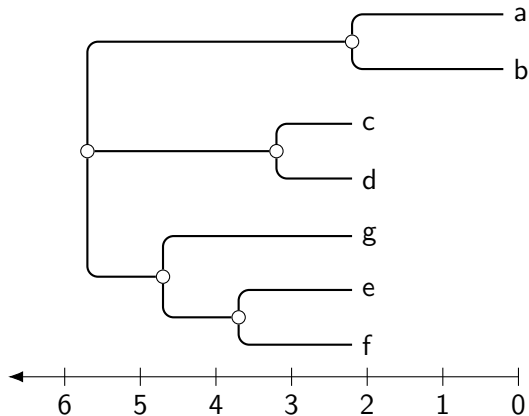
At what time did two observed genetic sequences coalesce?

- (We observe  $N$  genetic sequences)

Related problems:

- Infer mutation rates
- Discover variants of interest/variants of concern (clade emergence)
- Impute missing or ancestral sequences

# Phylogenetics



- Leaf nodes: Observed RNA sequences
- Interior nodes: Unobserved
- Goal: Infer tree topology and branch lengths

# Bayesian phylogenetics

- Place a prior on trees (Kingman's coalescent), develop proposals (generalized stepping-stone sampling), perform MCMC (BEAST)
  - Largest dataset studied with BEAST:  $\sim 25k$  taxa (L. Lyu et al. PNAS 2025)
- Construct a variational family, perform *variational Bayes* (vB: more scalable?)
  - vB is an iterative inference algorithm that approximates the posterior using a family of functions (in contrast, MCMC approximates with a set of samples)
- Note: These methods both require a likelihood function to link the tree topology and branch lengths to the observed sequences (JC, K2P, GTR)

# Single-linkage clustering

Construct a rooted tree topology from a distance matrix:

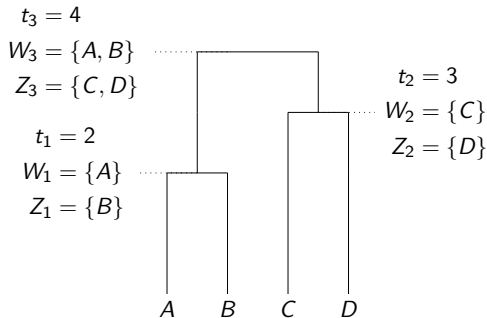
- 1: **Input:** Distances  $\mathbf{T} \in \mathbb{R}_{>0}^{\binom{N}{2}}$  and taxa set  $\mathcal{X} = \{\{x_1\}, \{x_2\}, \dots, \{x_N\}\}$ .
- 2: **for**  $n = 1, \dots, N - 1$  **do**
- 3:    $w^*, z^* \leftarrow \arg \min_{w, z} \{t^{\{w, z\}} : w, z \text{ have not coalesced by the } (n - 1)\text{-st event}\}$
- 4:   Set  $W_n \in \mathcal{X}$  to be the set containing  $w^*$
- 5:   Set  $Z_n \in \mathcal{X}$  to be the set containing  $z^*$
- 6:    $t_n \leftarrow t^{\{w^*, z^*\}}$
- 7:   Remove  $W_n, Z_n$  from  $\mathcal{X}$  and add  $W_n \cup Z_n$  to  $\mathcal{X}$
- 8: **end for**
- 9:  $\tau \leftarrow \{\{W_n, Z_n\}\}_{n=1}^{N-1}$
- 10:  $\mathbf{t} \leftarrow \{t_n\}_{n=1}^{N-1}$
- 11: **Return**  $(\tau, \mathbf{t})$

# Notation

- $\{W_n, Z_n\}$  is a bipartition (a.k.a. subsplit)
  - This means the  $n$ -th coalescent event involves a clade with leaves  $W_n$ , and a clade with leaves  $Z_n$ . ( $W_n, Z_n \subseteq \{x_1, \dots, x_N\}$ ,  $W_n \cap Z_n = \emptyset$ )
- $\{\{W_n, Z_n\}\}_{n=1}^{N-1}$  is a sequence of bipartitions
  - The topology of a tree with  $N$  leaves can be uniquely described by  $N - 1$  bipartitions
  - $\#W_1 = \#Z_1 = 1$ ,  $W_{N-1} \cup Z_{N-1} = \{x_1, \dots, x_N\}$
- For  $n \leq N - 1$ ,  $w$  and  $z$  have coalesced by the  $n$ -th event if there is a bipartition  $W_m, Z_m$  with  $w \in W_m, z \in Z_m$  and  $m \leq n$

## Single-linkage clustering (cont)

$$T = \begin{matrix} & A & B & C & D \\ A & * & \mathbf{2} & 8 & \mathbf{4} \\ B & * & * & 4.5 & 7 \\ C & * & * & * & \mathbf{3} \\ D & * & * & * & * \end{matrix}, \quad T = \begin{matrix} & A & B & C & D \\ A & * & \mathbf{2} & 5 & 6 \\ B & * & * & \mathbf{4} & 7 \\ C & * & * & * & \mathbf{3} \\ D & * & * & * & * \end{matrix},$$



Example: Two matrices  $T$  result in the same phylogenetic tree after running single-linkage clustering. Entries of  $T$  that trigger coalescence are bolded

## Related work: GeoPhy

- Associate  $i$ -th taxa with a point  $x_i$  in  $k$ -dimensional space
- The variational family is a distribution  $q$  on  $X \in \mathbb{R}^{N \times D}$
- Given a draw  $X$ , the corresponding tree is found by forming the distance matrix  $T$  such that  $t^{(i, j)}$  is the Euclidean distance between  $x_i$  and  $x_j$ , and then running single-linkage clustering
- It is difficult to find a closed form for the distribution on tree topologies induced by  $q$ , so the variational family is represented by samples
- The method is generalized to points on a hyper-sphere, with distances given by geodesics

— *T. Mimori, M. Hamada. NeurIPS 2023*



## Related work: VBPI

... Variational Bayesian Phylogenetic Inference

- Run an initializing MCMC in BEAST
- Record all subsplits (bipartitions) appearing in the initializing MCMC run
- Form subsplit Bayesian network (SBN) from these subsplits
- The variational family is a distribution on the SBNs
- Excellent accuracy in likelihood estimation
- May not be scalable, as number of subsplits in the MCMC run is large

— C. Zhang, F.A. Matsen IV. ICLR 2019

# VIPR: Variational inference with products . . .

. . . over bipartitions

- Like GeoPhy, we consider single-linkage clustering to map from distance matrices to a tree
- We place the variational family as a distribution directly on the distance matrix (each off diagonal entry of the upper triangle is log normal)
- We derive the distribution of a tree implied by the variational family
- This distribution has a closed form as a sum over bipartitions (subsplits)

— *E. Sidrow, A. Bouchard-Côté and L.T. Elliott. ICML 2025*

# The probability of a tree has a closed form ...

... in terms of the distribution of the distance matrix  $T$

**Proposition.** If the random variables  $t^{\{u,v\}}$  are mutually independent, and all  $q_\phi^{\{u,v\}}$  are continuous in  $\phi$  and  $t$  for all  $\{u, v\}$  with  $u, v \in \mathcal{X}$ , and  $Q_\phi^{\{u,v\}}$  is the survival function of  $t^{\{u,v\}}$ , then  $q_\phi(\tau, \mathbf{t})$  has the following form:

$$q_\phi(\tau, \mathbf{t}) = \prod_{n=1}^{N-1} \left( \left( \sum_{\substack{w \in W_n \\ z \in Z_n}} \frac{q_\phi^{\{w,z\}}(t_n)}{Q_\phi^{\{w,z\}}(t_n)} \right) \prod_{\substack{w \in W_n \\ z \in Z_n}} Q_\phi^{\{w,z\}}(t_n) \right).$$

Here  $W_n, Z_n$  is the bipartition induced by the  $n$ -th coalescent,  $\tau$  is the tree topology, and  $t$  are the coalescent times

## Intuition

$$\left( \sum_{\substack{w \in W_n \\ z \in Z_n}} \frac{q_{\phi}^{\{w,z\}}(t_n)}{Q_{\phi}^{\{w,z\}}(t_n)} \right) \prod_{\substack{w \in W_n \\ z \in Z_n}} Q_{\phi}^{\{w,z\}}(t_n) = \sum_{\substack{w \in W_n \\ z \in Z_n}} \left( q_{\phi}^{\{w,z\}}(t_n) \prod_{\substack{w' \in W_n \\ z' \in Z_n \\ \{w',z'\} \neq \{w,z\}}} Q_{\phi}^{\{w',z'\}}(t_n) \right)$$

- If two clades coalesce at time  $t_n$ , one pair (with one taxa from one clade, the other taxa from the other clade) must coalesce at time  $t_n$  in the distance matrix. And all other such pairs must coalesce after time  $t_n$  in the distance matrix (survival function  $Q(\cdot) = \Pr(rv \geq \cdot)$ ). This marginalizes which pair coalesced
- Since entries in the distance matrix are independent, the above marginalization is summed over coalescent events

# Complexity

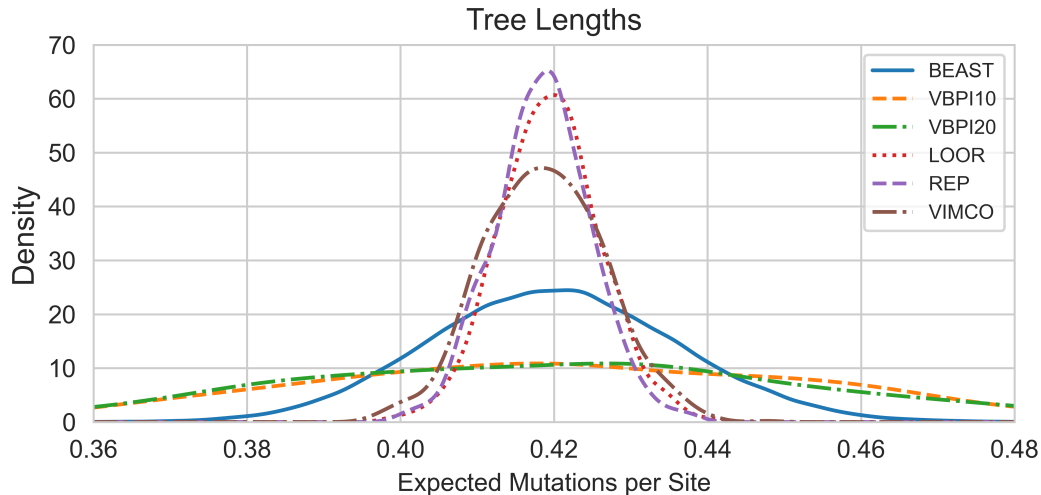
$$\prod_{n=1}^{N-1} \left( \left( \sum_{\substack{w \in W_n \\ z \in Z_n}} \frac{q_{\phi}^{\{w,z\}}(t_n)}{Q_{\phi}^{\{w,z\}}(t_n)} \right) \prod_{\substack{w \in W_n \\ z \in Z_n}} Q_{\phi}^{\{w,z\}}(t_n) \right).$$

- Each pair  $w, z$  occurs once in one of the inner sums, and once in one of the inner products:  $\mathcal{O}(N^2)$  operations. There are  $N - 1$  terms in total, so an additional  $\mathcal{O}(N)$  operations
- Total complexity for evaluating  $q$  or derivatives of  $q$ :  $\mathcal{O}(N^2 + N) = \mathcal{O}(N)$

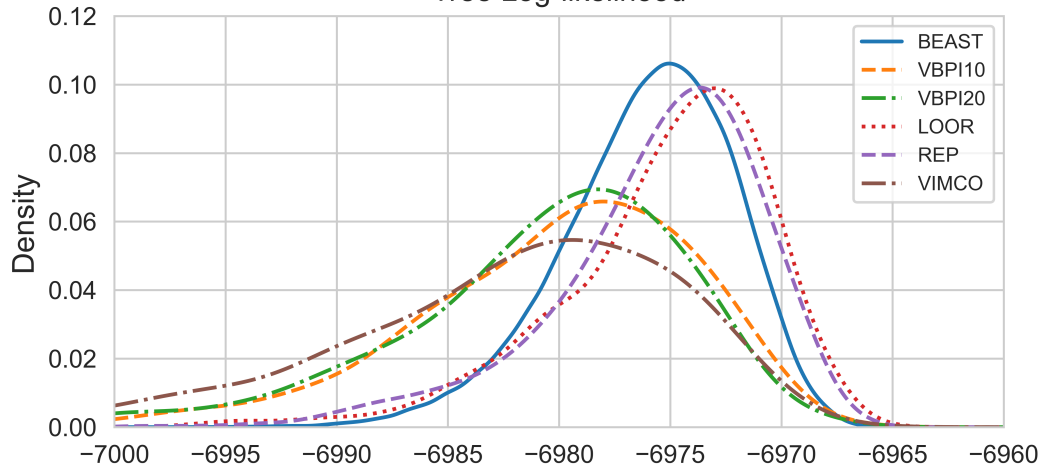
# Inference

- Derivatives of the sum product in the proposition are found using automatic differentiation (torch/autodiff)
- ELBO optimization is done using:
  - VIMCO (Mnih, Rezende 2016)
  - LOOR
  - or the reparameterization-trick

## Results: DS1 ( $N = 27, M = 1949$ )

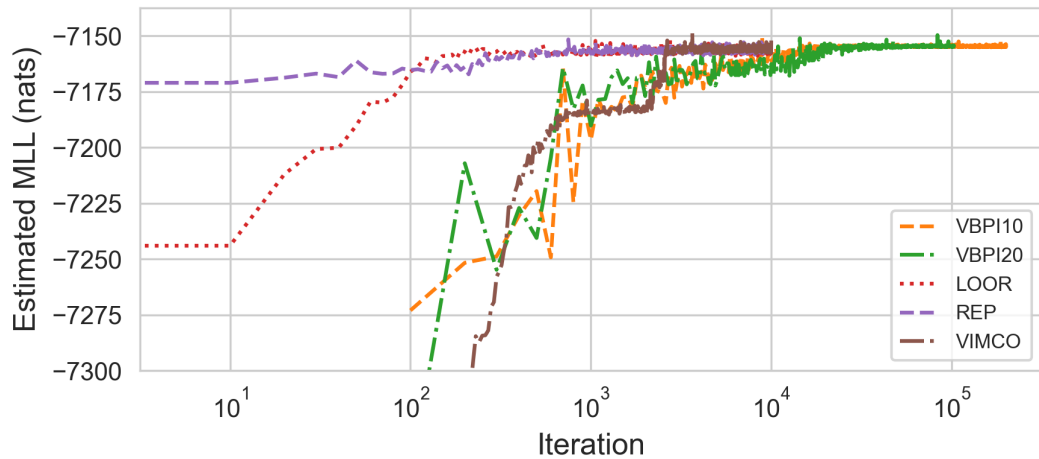


## Tree Log-likelihood



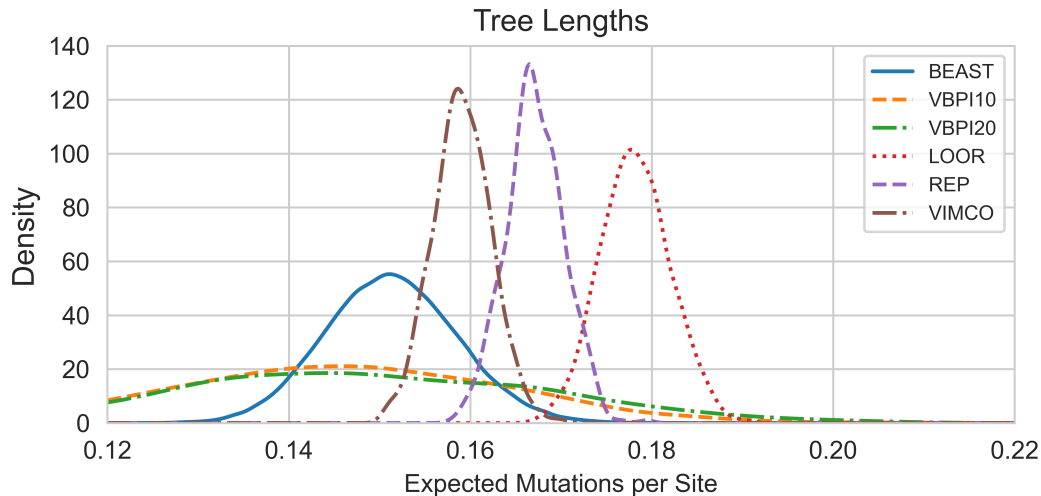


DS1

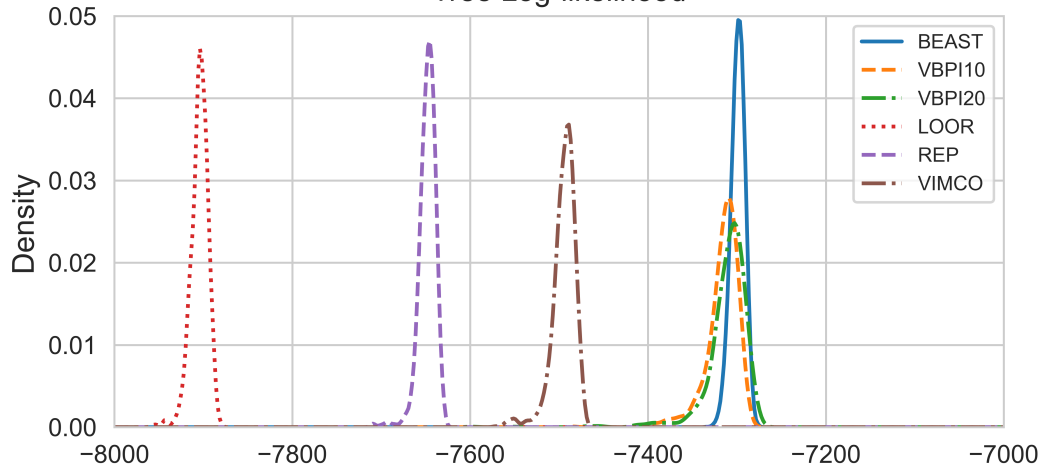


Thank You!

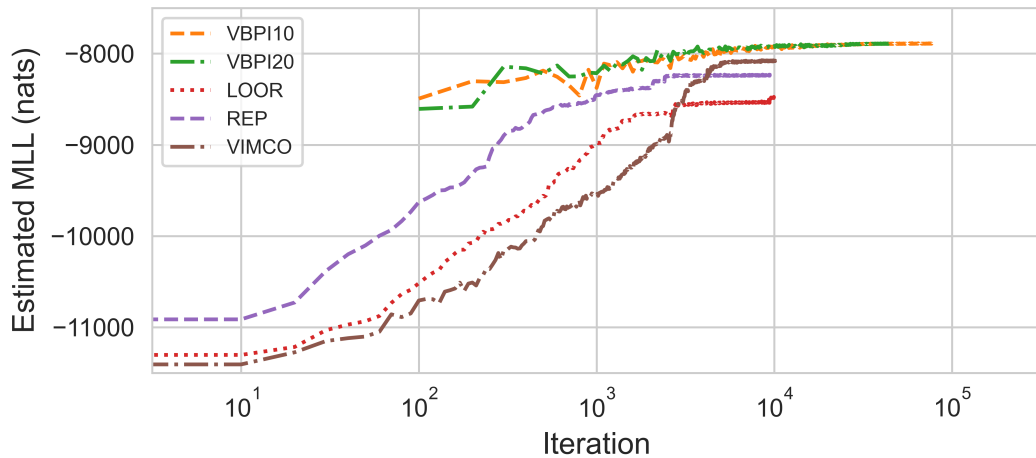
## Results: COV ( $N = 72, M = 3101$ )



Tree Log-likelihood



# COVID-19 Dataset



*Table 3. Number of tree structure parameters versus number of taxa (NTAXA) on simulated data with 1,000 sites.*

NTAXA	VBPI	VIPR
8	4	56
16	44	240
32	55	992
64	3,826	4,032
128	29,939	16,256
256	127,217	65,280
512	319,533	261,632